IPsec RoCEv2

Boris Pismenny

March 2018
What is RoCEv2?

- RoCEv2 – RDMA over Converged Ethernet ( Routable )
- RoCEv2 is a Supplement to InfiniBand Architecture Specification
- RoCEv2 is implemented in the RDMA subsystem in Linux ( ib_uverbs )
- RoCEv2 uses UDP destination port 4791
RoCEv2 => RoCEv2 + IPsec

- EtherType indicates that packet is RoCE (i.e. next header is IB GRH)
- IP header indicates that packet is IP (i.e. next header is IB BTH)
- UDP dport number indicates that next header is IB BTH
RoCEv2 UDP Ports

- In InfiniBand a Queue Pair (QP) is like a socket in Ethernet
  - A connection is formed between two QPs

- RDMA there are a few QP types:
  1. Reliable/Unreliable Connected (RC, UC) – RC QP is like TCP
  2. Reliable/Unreliable Datagram (RD, UD) – UD QP is like UDP

- UDP source port is constant for the duration of a connected QP (RC,UC)
  - For RD,UD each datagram may use a different UDP source port

- Is there a 1:1 mapping between 5-tuple and RDMA QP?
  - No… There could be more QPs than the number of UDP ports between two hosts:
    - Only $2^{16}$ UDP source ports
    - There could be up to $2^{24}$ QPs between two hosts
  - Moreover, UD QPs can choose the source port per datagram
QP Numbers and IPsec Selectors

- RoCEv2 packets are just UDP packets with an additional BTH header
  - BTH headers contain the destination QP number

- Hardware knows the source QP number while sending a packet and the destination QP when receiving packets

- We could use the source-QP/destination-QP number to form the outgoing/incoming IPsec policy
RoCEv2 IPsec API

**General idea:**
- Reuse the existing XFRM and IKE frameworks for the control path (just like sockets)
- Supported via RDMA Connection Manager (`rdma_cm`) or via Full Offload in IKE

**Two ways of configuring IPsec:**
1. Set per QP (like `IP_XFRM_POLICY` socket option)
2. Set full offload IPsec on UDP dport 4791 (could use IKE)
3. Manually
RoCEv2 IPsec API - Per QP

- Set a new rdma_cm option (just like a setsockopt)

- Add a xfrm state lookup for rdma_connect called rdma_xfrm_lookup
  - like xfrm_lookup, but using a QP instead of a socket
  - rdma_xfrm_lookup – finds full offload policy and creates a full offload xfrm_state for it

- Call km_query to establish a new SA
  - How to provide the QP numbers?

- New Transparent SA is established
  - QP connection establishment resumes (just like in XFRM with sockets)
RoCEv2 IPsec API – using IKE

- Basic configuration – set the IP addresses and RoCEv2 UDP port as the IKE policy with full IPsec offload
  - No support for QP number policies
  - New QP number selector type for IKEv2?
  - If IKEv2 supported QP numbers, then how would we add it to XFRM netlink?

- `rdma_xfrm_lookup` could check for a global matching IPsec policy and trigger `km_query`
Thank You